Journal of Infectious Diseases Advance Access published October 30, 2014

Tracking a tuberculosis outbreak over 21 years: strain-specific single nucleotide polymorphism-typing combined with targeted whole genome sequencing

David Stucki[1,2,§], Marie Ballif[3,§], Thomas Bodmer[4,5], Mireia Coscolla[1,2], Anne-Marie Maurer[6], Sara Droz[4], Christa Butz[7], Sonia Borrell[1,2], Christel Längle[4], Julia Feldmann[1,2], Hansjakob Furrer[8], Carlo Mordasini[7], Peter Helbling[9], Hans L. Rieder[10,11], Matthias Egger[3,12], Sébastien Gagneux[1,2,*], Lukas Fenner[1,2,3,*]

[1]Swiss Tropical and Public Health Institute, 4051 Basel, Switzerland

[2]University of Basel, 4003 Basel, Switzerland

[3]Institute of Social and Preventive Medicine, University of Bern, 3012 Bern, Switzerland

[4]Institute for Infectious Diseases, University of Bern, 3012 Bern, Switzerland

[5]labormedizinisches zentrum Dr Risch, 3097 Liebefeld-Bern, Switzerland

[6]Cantonal Health Authorities, 3011 Bern, Switzerland

[7]Bernese Lung Association, 3007 Bern, Switzerland

[8]Department of Infectious Diseases, Bern University Hospital and University of Bern, 3010 Bern, Switzerland

[9]Federal Office of Public Health, 3003 Bern, Switzerland

[10]Tuberculosis Department, International Union Against Tuberculosis and Lung Disease, 75006 Paris, France

[11]Institute of Social and Preventive Medicine, University of Zurich, 8001 Zurich, Switzerland

[12]School of Public Health and Family Medicine, University of Cape Town, 7925 Cape Town, South Africa

[*]Corresponding authors Dr. Lukas Fenner, Swiss Tropical and Public Health Institute, Basel, Switzerland, E-Mail: lukas.fenner@unibas.ch, Phone: +41-61-284-8369; Fax: +41-61-284-8101, Prof. Sebastien Gagneux, Swiss Tropical and Public Health Institute, Basel, Switzerland, E-Mail: sebastien.gagneux@unibas.ch, Phone: +41-61-284-8369; Fax: +41-61-284-8101

[§]Contributed equally

ABSTRACT

Background: Whole genome sequencing (WGS) is increasingly used in molecular-epidemiological investigations of bacterial pathogens, despite cost- and time-intensive analyses. We combined strain-specific single nucleotide polymorphism (SNP)-typing and targeted WGS to investigate a tuberculosis cluster spanning 21 years in Bern, Switzerland.

Methods: Based on genome sequences of three historical outbreak *Mycobacterium tuberculosis* isolates, we developed a strain-specific SNP-typing assay to identify further cases. We screened 1,642 patient isolates, and performed WGS on all identified cluster isolates. We extracted SNPs to construct genomic networks. Clinical and social data were retrospectively collected.

Results: We identified 68 patients associated with the outbreak strain. Most were diagnosed in 1991-1995, but cases were observed until 2011. Two thirds belonged to the homeless and substance abuser milieu. Targeted WGS revealed 133 variable SNP positions among outbreak isolates. Genomic network analyses suggested a single origin of the outbreak, with subsequent division into three sub-clusters. Isolates from patients with confirmed epidemiological links differed by 0-11 SNPs.

Conclusions: Strain-specific SNP-genotyping allowed rapid and inexpensive identification of *M. tuberculosis* outbreak isolates in a population-based strain collection. Subsequent targeted WGS provided detailed insights into transmission dynamics. This combined approach could be applied to track bacterial pathogens in real-time and at high resolution.

INTRODUCTION

Tuberculosis transmission has traditionally been investigated using contact tracing and molecular typing [1,2]. However, social contact data are often hard to obtain retrospectively, especially in high-risk groups such as homeless and substance abusers who are difficult to trace [3–7]. Moreover, classical molecular epidemiological techniques like *IS*6110 restriction fragment length polymorphism (RFLP) and mycobacterial interspersed repetitive unit – variable number of tandem repeat (MIRU-VNTR) interrogate only a small proportion of the mycobacterial genome and therefore suffer from limited resolution [8].

More recently, whole genome sequencing (WGS) of *Mycobacterium tuberculosis* has been used to investigate tuberculosis outbreaks [9]. Known as “genomic epidemiology” [10], this emerging field uses WGS to detect unknown transmission events, identify ”super-spreaders”, and exclude or confirm epidemiologically suspected transmission links [11–14]. Moreover, WGS can also be used to detect drug resistance mutations [15]. Even though routine WGS has the potential to replace classical genotyping [14,16], analysing WGS data remains resource-intensive, and requires further standardization to meet public health needs, in particular for tracking ongoing outbreaks in real-time [9].

In 1993, a tuberculosis outbreak was reported in the Canton of Bern, Switzerland [17]. Twenty-two cases were involved, and their *M. tuberculosis* isolates shared identical *IS*6110 RFLP patterns. As in other affluent countries [18,19], this outbreak involved mainly homeless and substance abusers. In 2012, we studied the molecular epidemiology of tuberculosis in Switzerland. We used the classical methods spoligotyping and MIRU-VNTR to genotype 520 *M. tuberculosis* isolates from patients diagnosed with tuberculosis between 2000 and 2008 (12.3% of all culture-

confirmed tuberculosis cases in Switzerland during the study period) [20]. Among 68 isolates from the Canton of Bern, two were identified as belonging to the same outbreak described in 1993, indicating that this particular strain was still circulating in the region.

In the present study, we used a combination of single-nucleotide polymorphism (SNP) typing and targeted WGS to track the spread of the outbreak over two decades. Using representative isolates from the historical outbreak, we first developed a novel strain-specific SNP-typing assay to rapidly and inexpensively identify all tuberculosis cases caused by this strain in the Canton of Bern between 1991 and 2011. We then applied targeted WGS to all cluster isolates identified by the screening assay to study the outbreak dynamics in relation to social contact information.

METHODS

Study setting and sample set

We subcultured 1,642 patient isolates available from the *M. tuberculosis* strain collection at the Institute for Infectious Diseases, Bern, Switzerland. These isolates were all collected between 1991 and 2011, and correspond to 84.6% of all 1,940 tuberculosis cases (all forms) notified in the Canton of Bern during the same time period (Figure 1). Subcultures were performed on Löwenstein-Jensen slants according to international laboratory standards. Purified DNA for WGS was obtained using the CTAB extraction method after subculturing a single colony in 7H9 liquid medium [21]. Bulk extracts (i.e. without single colony selection) were available for four isolates.

The collection included the 22 historical outbreak patient isolates reported by Genewein et al. [17]. One strain isolated in 1987, before the systematic collection was started in 1991, was also included (P028, originally labeled as patient "1" in [17]). Finally, for one outbreak patient from 1991, we identified an additional strain isolated in 1988 (P006A, originally patient "2", Figure 1).

Cluster strain-specific SNP-typing assay and screening of strain collection

We performed WGS on one historical outbreak isolate from 1992 [17] and two isolates from 2005 and 2008 with the same "Bernese cluster" MIRU-VNTR and spoligotyping patterns [20], as described below. We also performed WGS on two isolates with the same spoligotyping pattern (isolation 2001 and 2004), but a different MIRU-VNTR pattern (3 and 4 loci different compared to the outbreak isolate), one additional Lineage 4 isolate from another study, and the reference strain H37Rv (see Figure 2). The three outbreak isolates shared 118 SNPs not observed in any of the control isolates (Figure 2). We used one of these outbreak-specific SNPs (878,174 GA, position in reference to H37Rv) to develop a real-time PCR SNP-typing assay (TaqMan, Life Technologies, Switzerland), as described previously [22]. All 1,642 available isolates were screened for the presence of that SNP. For confirmation, we subjected all isolates with a mutation at this position to a second, phylogenetically redundant, SNP-typing assay (981,565 CT). Both SNPs were selected to be synonymous and located in genomic regions suitable for primer and probe design.

Whole genome sequencing and phylogenetic analyses

All isolates identified by the screening assay and the additional serial isolate from patient P006 were subjected to Illumina WGS at GATC Biotech Company (Konstanz,

Germany) with a median nucleotide coverage of 157.3 reads (range 29.1-896.9). Sequence read mapping and SNP-calling was done as previously described [23]. We considered SNPs with a coverage of at least 10 sequencing reads and a value of 20 in the phred-scaled quality score. SNPs in genes annotated as “PE/PPE/PGRS”, “maturase”, “phage”, “insertion sequence” or “13E12 repeat family protein” were removed. Additionally, positions with missing nucleotide calls in at least three isolates were excluded. We used a second short-read alignment tool (SMALT, Wellcome Trust Sanger Institute, UK) to obtain SNP calls. Only positions called by both methods after filtering for the above-mentioned criteria were included for further analysis. A subset of 28 SNPs was confirmed by Sanger sequencing (Supplementary Information).

A genomic network with all variable positions was generated using Fluxus Network Software (www.fluxus-engineering.com) and the Median Joining Algorithm. Arlequin 3.5.1.12 [24] was used to calculate genetic distances and fixation indices (F_{ST}) to estimate population separation between genomic sub-clusters. Statistical significance was calculated with permutations.

Raw sequencing data are available under accession number PRJEB5925 (European Nucleotide Archive).

Clinical and socio-demographic data collection

We collected clinical and socio-demographic data on all patients identified as belonging to the cluster. Treating physicians, hospital archives, the Bernese Lung Association, and the Cantonal health authorities collected the data using standardized questionnaires. We also collected contact tracing information for confirmed or presumptive links among cluster patients. The National Tuberculosis

Surveillance Registry (Federal Office of Public Health) provided basic demographic data (age, sex, birth place, disease site) for all tuberculosis cases notified in the Canton of Bern between 1991 and 2011.

Definitions

We categorized patients into new tuberculosis cases, recurrent cases or unknown previous treatment status according to international definitions [25]. *Confirmed* links between cases were defined for contacts named in the contact tracing information. *Presumptive* links between cases were defined when contacts were not clearly named, but strongly supported by other contact tracing information (visiting common hotspots of transmission, shared housing, shared place of work). Alcohol abuse was defined as daily consumption of alcohol, and smoking as past or current smoking. We defined "milieu" as a combined variable capturing high-risk populations (substance abusers and/or homeless individuals), and/or patients frequenting high-risk settings (i.e. drug injection places, methadone distribution places, homeless shelters).

Statistical analyses

We used $\chi 2$ tests or Fisher's exact tests to assess differences between groups in binary variables and the Wilcoxon rank sum test for continuous variables. We investigated i) differences between the characteristics of "Bernese cluster" patients and all other notified tuberculosis cases in the Canton of Bern between 1991 and 2011, and ii) differences between patients in the genomic sub-clusters.

Ethics statement

Ethics approval for this study was obtained from the Ethics Committee of the Canton of Bern, Switzerland. The treating physicians sought written informed consent from study participants. In most cases, informed consent could, however, not be obtained because the patient could not be located or was known to have died. We therefore obtained permission from the Federal Expert Commission on Confidentiality in Medical Research to use the data provided by the treating physicians.

Role of the funding sources

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

RESULTS

Identification of "Bernese cluster" isolates by strain-specific SNP-typing

Using the strain-specific real-time PCR SNP-typing assay, we screened 1,642 *M. tuberculosis* single patient isolates for the "Bernese cluster"-specific SNP (878,174 GA) and identified 71/1,642 (4.3%) isolates as belonging to this cluster. All isolates except for three were confirmed using a second, phylogenetically redundant SNP (981,565 CT). These three isolates with ambiguous results were excluded from further analysis, as subsequent WGS revealed mean pairwise distances of 91, 148, and 174 SNPs to the other cluster isolates, respectively. In contrast, all other cluster isolates were separated by 19 SNPs or less (Supplementary Table 1). This corresponds to a specificity of 99.8% (3 false-positive out of 1,574 non-cluster

isolates), when considering our screening results based on only the first strain-specific SNP. All 22 historical isolates described in 1993 [17] were correctly identified as "Bernese cluster" genotype by both SNP-typing assays (22/22, sensitivity 100%). Hence, we identified a total of 68 patients linked to the cluster strain (Figures 1 and 3).

For one patient, two isolates were available, and we included both for further analyses (P006A and P006B, isolated in 1988 and 1991, respectively) because of the central role of this patient described in the original study (originally labeled patient "2" [17]). Hence, we included a total of 69 cluster isolates in the WGS analyses (Figure 1).

To illustrate the strengths of our novel combined approach, we compared the costs, advantages and disadvantages of the different methods used in outbreak investigations of tuberculosis (Table 1). Our strain-specific SNP assay to screen 1,642 isolates was approximately 15 times less expensive than the current gold standard based on MIRU-VNTR. A combination of the SNP-assay and targeted WGS for 69 cluster isolates was approximately 20 times less expensive than performing WGS of the entire collection.

Description of the tuberculosis cluster over 21 years

Of the 68 "Bernese cluster" patients, 55 (80.9%) were diagnosed until 1998, and 13 were diagnosed between 1999 and 2011 (Figure 3). The characteristics of the 68 cluster patients compared to all other tuberculosis cases diagnosed in the same region and time period are presented in Table 2. Cluster patients were more likely to be male (79.4% vs 57.3%, p<0·001), born in Switzerland (83.8% vs 47.4%, p<0.001), and to have pulmonary tuberculosis as opposed to extrapulmonary tuberculosis

(94.1% vs 75.1%, p<0.001). The median age of the cluster patients was 41 years (IQR 34 -53), versus 44 years (29-71) for all other tuberculosis patients (p=0.12). Most of the cluster patients were part of the local injection drug scene and/or homeless milieu (67.6%). Among cluster patients, 19.1% were HIV-infected (HIV information was unavailable for non-cluster patients). Four particular hotspots of tuberculosis transmission were identified within the milieu: one short-term homeless shelter, one long-term social integration home, one meeting point for injection drug and methadone supply, and a bar where substance abusers met. The distribution of cluster patients among these four transmission hotspots, the social milieu, and the general population is presented in Figure 4.

Contact investigation provided information on potential patient-to-patient links (Figure 4). Fourteen of 68 patients (20.6%) had *confirmed* epidemiological links (red lines in Figure 4). Confirmed links were more frequent between patients sharing transmission hotspots, indicating the large degree of social interaction in these settings. Only one confirmed link was identified in the general population, namely between a father and his daughter (P040-P041).

Whole genome sequencing of "Bernese cluster" isolates

A total of 133 variable positions were identified among the 69 cluster isolates (Supplementary Table 2). We generated a Median Joining network using these 133 variable positions (Figure 5). Despite identical MIRU-VNTR and spoligotyping patterns (Supplementary Table 3), 52/69 isolates (75.4 %) were discriminated by at least one SNP from their most closely related neighbor. The maximum number of SNPs between any two isolates was 19 (Supplementary Table 1), and the mean pairwise distance between all isolates was 6.0 SNPs (standard deviation 2.9). Patient

isolates with confirmed epidemiological links differed by between zero and 11 SNPs (Supplementary Table 1). No drug-resistance associated mutation was detected among the 69 cluster isolates (Supplementary Information).

Sub-clusters and key patient

The genomic network revealed three independent star-like structures, suggesting i) an early divergence of a shared ancestor strain into three sub-clusters (1-3 in Panel A of Figure 5), and ii) the presence of one or several "super-spreaders". A fourth, more distantly related sub-cluster was separated by nine SNPs from the nearest isolate (P006A) (Panel A, Figure 5). In contrast, sub-clusters 1 and 2 were separated by one SNP, sub-clusters 1 and 3 by two SNPs. The average pairwise distance within each sub-cluster was 4.2 (sub-cluster 1), 3.1 (sub-cluster 2), 5 (sub-cluster 3), and 9.3 SNPs (sub-cluster 4). All corresponding SNP-distances were larger *between* sub-clusters than *within* sub-clusters (4.8 SNPs between sub-clusters 1 and 2, 6.8 SNPs between sub-clusters 1 and 3, 12.2 SNPs between sub-clusters 1 and 4, 7.3 SNPs between sub-clusters 2 and 3, 12.6 SNPs between sub-clusters 2 and 4, 14.6 SNPs between sub-clusters 3 and 4). Pairwise fixation indices (F_{ST}) between sub-clusters were between 0.24 (sub-cluster 1 and 2) and 0.67 (sub-cluster 2 and 4) (all $p<0.005$), further supporting the sub-cluster distinction.

The central position of sub-cluster 1 was occupied by the first isolate of patient P006 (P006A, isolated 1988) together with isolate P010. The second isolate of the key patient (P006B, isolated 1991) was found in the central position of sub-cluster 2. This suggests a second tuberculosis episode of the key patient P006, generating further secondary cases. Seven other isolates were genomically clustered with P006B and could also be the source of transmission in sub-cluster 2. However, patient P006 was

a homeless substance abuser known to have interrupted treatment, with a history of treatment failure, and a key role in the outbreak was already suspected in 1993.

The central position of sub-cluster 3 remained unoccupied (i.e. no isolate was available with this hypothetical genotype). This could be explained by an unsampled strain variant of patient P002 (i.e. from a mixed infection of micro-evolved strains) that had been transmitted to other sub-cluster 3 patients. Such a variant would have been missed as a result of the single colony isolation step preceding WGS. Manual inspection of the corresponding Sanger sequence trace files generated from a separate DNA preparation without single colony isolation step revealed a double peak at the SNP separating P002 from the hypothetical node mv2 (genomic position 2,156,041) in the bulk isolate of P002, indicating the presence of a mixed population in P002. Alternatively, an unsampled patient isolate (reported as “A”, “B”, “C” by Genewein [17]) might correspond to the central position of sub-cluster 3.

When plotting the time period of isolation of *M. tuberculosis* strains in the genomic network (Panel B, Figure 5), we found no evidence that the different sub-clusters were associated with specific time periods.

Finally, we compared the patient characteristics between the sub-clusters, excluding sub-cluster 4 (genetically distant, epidemiologically unrelated) and excluding patient P006, whose isolates belonged to both sub-clusters 1 and 2. We found that HIV infection was more frequent in sub-cluster 1 (7/17, 41.2%) than in the other sub-clusters (3/30, 9.7% in sub-cluster 2; 3/17, 17.6% in sub-cluster 3; p=0.04). All three sub-clusters included a majority of individuals from the social milieu (12/17, 70.6% in sub-cluster 1 [17/30 56.7%] in sub-cluster 2; 14/17 [82.4%] in sub-cluster 3; p=0.19). Sub-cluster 3 was associated with two particular transmission hotspots: a long-term social integration home (1/17 [5.9%] in sub-cluster 1; 0/30 in sub-cluster 2; 7/17

[41.2%] in sub-cluster 3; p<0.001; Panel C in Figure 5), and a meeting point for methadone supply (1/17 [5.9%] in sub-cluster 1; 1/30 [3.3%] in sub-cluster 2; 5/17 [29.4%] in sub-cluster 3; p=0.02).

DISCUSSION

Using a novel combination of a rapid and inexpensive strain-specific SNP screening assay and targeted WGS, we were able to track a tuberculosis cluster spanning 21 years, and to reveal the transmission dynamics among the outbreak patients.

Our study demonstrated the feasibility and advantages of tracking a tuberculosis outbreak using a strain-specific SNP-based screening assay in a large population-based collection of *M. tuberculosis* isolates. We subsequently performed targeted WGS on the 69 thus identified cluster isolates, which - combined with social contact data - enabled to retrace transmission dynamics at high resolution. The combined cost of six initial whole genome sequences used to design the SNP-typing assay and the subsequent population-wide screening by real-time PCR was low (approximately US$ 4,900), compared to the cost of screening all isolates with any other genotyping method. Additionally, the time required for screening was substantially reduced, making this a powerful approach for identifying and tracking of tuberculosis outbreaks in real-time. Even though our study was retrospective, our approach could be used to screen isolates prospectively.

Our results indicate that the sensitivity and specificity are nearly 100%. However, we could only estimate the technical test characteristic as WGS data were not available for the entire collection. Importantly, the performance of a strain-specific SNP-typing assay depends on the selection of SNPs, and the selection of SNPs depends on the

isolates initially sequenced. For the successful design of such an assay, we recommend the following: i) select at least two known clustered isolates for WGS, ii) include at least two control isolates with closely related but different genotyping patterns than the clustered isolates (e.g. MIRU-VNTR), iii) identify SNPs specific to all clustered isolates and absent in the control isolates, iv) exclude SNPs in genes known to be associated with drug resistance, v) select SNPs suitable for probe and primer design, vi) use at least two SNPs for the screening of isolates, as the specificity of each SNP might vary.

Applying our combined approach and linking it to clinical and contact tracing data, we found that the tuberculosis cluster continued to propagate in Bern, mainly in particular transmission hotspots and in the originally described high-risk populations of substance abusers and homeless people [17]. This is consistent with previous reports from other low-incidence settings [4,19,26,27]. The outbreak involved a key patient who caused numerous secondary cases, which corresponds to a “super-spreader” behavior. Most outbreak cases occurred between 1991 and 1995, followed by sixteen years of sporadic cases likely reflecting a majority of reactivation disease. The cases in the early 1990s coincided with known peaks of heroin abuse in Switzerland. However, 32.4% of all cluster cases involved “non-milieu” population, possibly reflecting transmission from the “milieu” to the wider community. In retrospect, more secondary cases could have been identified if our novel screening methodology had been available in the 1990s. Indeed, strain-specific SNP-typing would have provided an inexpensive method to identify outbreak cases more rapidly. Furthermore, targeted WGS would have identified “super-spreaders”, in a context where contact tracing is particularly difficult. Such “super-spreader” behavior could have then been targeted with intensified control measures to interrupt transmission.

The targeted WGS analyses of all cluster isolates identified by strain-specific SNP-typing shed new light on the outbreak transmission dynamics compared to traditional genotyping methods. Whereas MIRU-VNTR and spoligotyping showed identical genotyping patterns, WGS revealed distinct genotypes for 76.5% of the “Bernese cluster” isolates. In particular, we identified four genomic sub-clusters not revealed by classical genotyping, likely reflecting concomitant but independent clusters of a common ancestral strain. However, the genetic distances between sub-clusters were small (one, two, and three SNPs between sub-clusters 1, 2 and 3), and therefore, the definition of sub-cluster may be debatable. The sub-clusters were, however, supported by F_{ST} values indicating separation of these populations.

Two sequential isolates of the key patient (isolated in 1988 and 1991) occupied the central positions of sub-clusters 1 and 3, respectively. This suggests that two disease episodes of this patient lead to two independent star-like patterns in the genomic network, indicating a “super-spreader” behavior [12–14]. The central role of this patient was already suspected in the original description of the outbreak [17]. Hence, WGS analyses indicated that this patient likely caused more secondary cases than previously assumed.

Despite the many advantages of WGS, our results also showed that interpreting WGS data has limitations. For example, nearly 25% of cluster isolates were genomically indistinguishable from at least one other isolate. This emphasizes the need to include repetitive regions of the genome that are currently excluded due to technical limitations [28,29]. Furthermore, there is increasing evidence that bacterial populations within patients are heterogeneous as a consequence of ongoing micro-evolution, further complicating the interpretation of transmission events [30]. In our study, genomes were generated from single colonies for most isolates. Yet,

considering potential clonal variants that were randomly excluded from the sequencing process could influence the way transmission events are inferred. With improving sequencing technologies, future studies should sequence “bulk” isolates rather than single colonies, and consider within-host heterogeneity in bacterial populations. Mutations can also arise during laboratory culture; these could be avoided by performing WGS directly from sputum [31].

In conclusion, our strain-specific SNP-based screening approach offers a rapid and inexpensive way of tracking tuberculosis outbreaks retrospectively and prospectively. This novel screening method combined with targeted WGS can be used to guide control interventions by rapid and inexpensive screening, revealing transmission hotspots and missing links in transmission chains. Future studies could use this approach in real-time to track ongoing outbreaks of tuberculosis and other infectious diseases in hospital settings as well as population-wide.

ACKNOWLEDGMENTS

We thank all tuberculosis patients for participating in this study, the treating physicians and hospitals, as well as the Swiss HIV Cohort Study for providing clinical information, and the Institute for Infectious Diseases, University of Bern, Switzerland, for providing the clinical isolates. We are indebted to the National Tuberculosis Surveillance Registry at the Federal Office of Public Health, the Bernese Lung Association and the Cantonal health authorities for supporting the collection of clinical data and contact tracing information.

FUNDING

This work was supported by a grant from the Bernese Lung Association (Bern, Switzerland), the Swiss National Science Foundation (grant no. PP00P3_150750, grant number 33CS30_134277, Swiss HIV Cohort Study), the United States of America National Institutes of Health (grant no. AI090928, U01AI069924), and the European Research Council (grant no. 309540-EVODRTB).

CONFLICT OF INTERESTS

All authors declare no conflict of interest.

Meetings where the information has been presented

Parts of this work were presented at the following conferences:

- 10th International Meeting on Microbial Epidemiological Markers (IMMEM-10), Paris, October 2-5, 2013; Title "Genomic epidemiology of a tuberculosis outbreak in Switzerland over 21 years"; Abstract no. 69.
- The Union Congress, Paris, October 30-November 3, 2013; Title “Genomics meets public health: insights from a tuberculosis outbreak in Switzerland over 21 years”; Abstract no. OP-140-01.

REFERENCES

1. Cook VJ, Shah L, Gardy J, Bourgeois A-C. Recommendations on modern contact investigation methods for enhancing tuberculosis control. Int J Tuberc Lung Dis **2012**; 16:297–305.

2. McElroy PD, Rothenberg RB, Varghese R, et al. A network-informed approach to investigating a tuberculosis outbreak: implications for enhancing contact investigations. Int J Tuberc Lung Dis **2003**; 7:S486–493.

3. Cook VJ, Sun SJ, Tapia J, et al. Transmission network analysis in tuberculosis contact investigations. J Infect Dis **2007**; 196:1517–1527.

4. Anderson LF, Tamne S, Brown T, et al. Transmission of multidrug-resistant tuberculosis in the UK: a cross-sectional molecular and epidemiological study of clustering and contact tracing. Lancet Infect Dis **2014**; 14:406–415.

5. Asghar RJ, Patlan DE, Miner MC, et al. Limited utility of name-based tuberculosis contact investigations among persons using illicit drugs: results of an outbreak investigation. J Urban Health **2009**; 86:776–780.

6. Burki T. Tackling tuberculosis in London's homeless population. Lancet **2010**; 376:2055–2056.

7. Lambregts-van Weezenbeek CSB, Sebek MMGG, van Gerven PJHJ, et al. Tuberculosis contact investigation and DNA fingerprint surveillance in The Netherlands: 6 years' experience with nation-wide cluster feedback and cluster monitoring. Int J Tuberc Lung Dis **2003**; 7:S463–470.

8. Kato-Maeda M, Metcalfe JZ, Flores L. Genotyping of *Mycobacterium tuberculosis*: application in epidemiologic studies. Future Microbiol **2011**; 6:203–216.

9. Walker TM, Monk P, Smith EG, Peto TEA. Contact investigations for outbreaks of *Mycobacterium tuberculosis*: advances through whole genome sequencing. Clin Microbiol Infect **2013**; 19:796–802.

10. Gardy JL. Investigation of disease outbreaks with genome sequencing. Lancet Infect Dis **2013**; 13:101–102.

11. Bryant JM, Schürch AC, van Deutekom H, et al. Inferring patient to patient transmission of *Mycobacterium tuberculosis* from whole genome sequencing data. BMC Infect Dis **2013**; 13:110.

12. Gardy JL, Johnston JC, Sui SJH, et al. Whole-genome sequencing and social-network analysis of a tuberculosis outbreak. N Engl J Med **2011**; 364:730–739.

13. Roetzer A, Diel R, Kohl TA, et al. Whole genome sequencing versus traditional genotyping for investigation of a *Mycobacterium tuberculosis* outbreak: a longitudinal molecular epidemiological study. PLoS Med **2013**; 10:e1001387.

14. Walker TM, Ip CL, Harrell RH, et al. Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. Lancet Infect Dis **2013**; 13:137–146.

15. Köser CU, Bryant JM, Becq J, et al. Whole-genome sequencing for rapid susceptibility testing of *M. tuberculosis*. N Engl J Med **2013**; 369:290–292.

16. Diep BA. Use of whole-genome sequencing for outbreak investigations. Lancet Infect Dis **2013**; 13:99–101.

17. Genewein A, Telenti A, Bernasconi C, et al. Molecular approach to identifying route of transmission of tuberculosis in the community. Lancet **1993**; 342:841–844.

18. Abubakar I, Stagg HR, Cohen T, et al. Controversies and unresolved issues in tuberculosis prevention and control: a low-burden-country perspective. J Infect Dis **2012**; 205:S293–300.

19. Bamrah S, Yelk Woodruff RS, Powell K, Ghosh S, Kammerer JS, Haddad MB. Tuberculosis among the homeless, United States, 1994-2010. Int J Tuberc Lung Dis **2013**; 17:1414–1419.

20. Fenner L, Gagneux S, Helbling P, et al. *Mycobacterium tuberculosis* transmission in a country with low tuberculosis incidence: role of immigration and HIV infection. J Clin Microbiol **2012**; 50:388–395.

21. van Embden JD, Cave MD, Crawford JT, et al. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J Clin Microbiol **1993**; 31:406–409.

22. Stucki D, Malla B, Hostettler S, et al. Two new rapid SNP-typing methods for classifying *Mycobacterium tuberculosis* complex into the main phylogenetic lineages. PLoS ONE **2012**; 7:e41253.

23. Coscolla M, Lewin A, Metzger S, et al. Novel *Mycobacterium tuberculosis* complex isolate from a wild chimpanzee. Emerg Infect Dis **2013**; 19:969–976.

24. Excoffier L, Laval G, Schneider S. Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinform Online **2005**; 1:47–50.

25. Rieder HL, Watson JM, Raviglione MC, et al. Surveillance of tuberculosis in Europe. Working Group of the World Health Organization (WHO) and the European Region of the International Union Against Tuberculosis and Lung Disease (IUATLD) for uniform reporting on tuberculosis cases. Eur Respir J **1996**; 9:1097–1104.

26. Mitruka K, Oeltmann JE, Ijaz K, Haddad MB. Tuberculosis outbreak investigations in the United States, 2002-2008. Emerg Infect Dis **2011**; 17:425–431.

27. Zenner D, Southern J, van Hest R, et al. Active case finding for tuberculosis among high-risk groups in low-incidence countries. Int J Tuberc Lung Dis **2013**; 17:573–582.

28. Bryant JM, Harris SR, Parkhill J, et al. Whole-genome sequencing to establish relapse or re-infection with *Mycobacterium tuberculosis*: a retrospective observational study. Lancet Respir Med **2013**; 1:786–792.

29. Copin R, Coscollá M, Seiffert SN, et al. Sequence diversity in the pe_pgrs genes of *Mycobacterium tuberculosis* is independent of human T cell recognition. MBio **2014**; 5:e00960–00913.

30. Pérez-Lago L, Comas I, Navarro Y, et al. Whole genome sequencing analysis of intrapatient microevolution in *Mycobacterium tuberculosis*: potential impact on the inference of tuberculosis transmission. J Infect Dis **2014**; 209:98–108.

31. Blainey PC. The future is now: single-cell genomics of bacteria and archaea. FEMS Microbiol Rev **2013**; 37:10.1111/1574–6976.12015.

TABLES

Table 1. Comparison of tuberculosis outbreak tracking methods, considering the scenario of the present study (68 outbreak patients out of 1,642 patient isolates to be screened).

	This study		WGS of entire collection	MIRU-VNTR	Spoligotyping	Contact tracing
	SNP-assay for identification of cluster isolates (n=1,642)	Targeted WGS of identified cluster isolates (n=69)	(n=1,642)	(n=1,642)	(n=1,642)	
Estimated costs (USD)[1]						
per isolate	3 (SNP assay)	330 (targeted WGS)	330	49	26	High
total	4,926 [2] (SNP assay)	22,770 (targeted WGS)	541,860	80,458	42,692	
	Total: 27,696					
Advantages	Rapid identification of cluster isolates; inexpensive	Highest resolution among cluster isolates; additional information (e.g. drug-resistance mutations)	Highest resolution among all isolates (in all clusters); additional information obtained (e.g. drug-resistance mutations)	Current gold-standard for molecular epidemiology; can be semi-automatized	Low technology requirement	Information on transmission hotspots for targeted prevention; information on secondary cases
Disadvantages	No further resolution among cluster isolates; performance of assay depends on selection of initial isolates for WGS (e.g. SNP-selection)	Previous identification of outbreak isolates necessary	Expensive; extensive bioinformatic expertise necessary	Limited resolution within outbreak clusters	Low resolution of outbreak cluster analysis when used as single method	Expensive and time-consuming; misses many cases, particularly in high-risk populations
Prospective use	Can be used in real-time once outbreak is identified and an assay established. If used in combination with targeted WGS: highest resolution among outbreak isolates	In combination with SNP-assay	Can be used in real-time once bioinformatics expertise is established	Routine use	Yes	Yes

MIRU-VNTR, mycobacterial interspersed repetitive unit – variable number of tandem repeat; SNP, single nucleotide polymorphism; WGS, whole genome sequencing

[1] Cost calculations were based on commercially available services (www.genoscreen.fr; www.gatc-biotech.com), or estimated according to in-house costs (as of August 2014)

[2] WGS of six initial isolates, and screening using the strain-specific SNP-genotyping assay of 1,642 isolates

Table 2. Patient characteristics of the confirmed tuberculosis cluster cases compared to all other notified tuberculosis cases in the Canton of Bern between 1991 and 2011. Not all patient characteristics were available for the notified tuberculosis cases.

Characteristic	"Bernese cluster" TB cases	All other TB cases	P-value
	n (%)	n (%)	
Total	*68 (100)*	*1872 (100)*	
Age at TB diagnosis (median, IQR)	41 (34-53)	44 (29-71)	0.12
Sex			<0.001
Male	54 (79.4)	1072 (57.3)	
Female	14 (20.6)	800 (42.7)	
Birth region			<0.001
Switzerland	57 (83.8)	888 (47.4)	
Europe (without Switzerland)	10 (14.7)	398 (21.3)	
Sub-Saharan Africa	0	227 (12.1)	
Asia	1 (1.5)	284 (15.2)	
Caribbean and Latin America	0	31 (1.7)	
Other regions	0	36 (1.9)	
Unknown	0	8 (0.4)	
TB disease site			<0.001
Pulmonary	64 (94.1)	1406 (75.1)	
Extra-pulmonary	4 (5.9)	466 (24.9)	
TB category			
New case	54 (79.4)		
Recurrent	7 (10.3)		
Unknown	7 (10.3)		
Imprisonment within 2 years of diagnosis	9 (13.2)		
Diabetes	3 (4.4)		
Alcohol abuse	39 (57.4)		
Smoker	41 (60.3)		
Injection drug user	18 (26.5)		
Homeless	21 (30.9)		
HIV positive	13 (19.1)		
Homeless/substance abuser milieu	46 (67.6)		
Residence			
Bern City	37 (54.4)		
Outside Bern City	29 (42.6)		
Unknown	2 (2.9)		

TB, tuberculosis; IQR, interquartile range

FIGURES

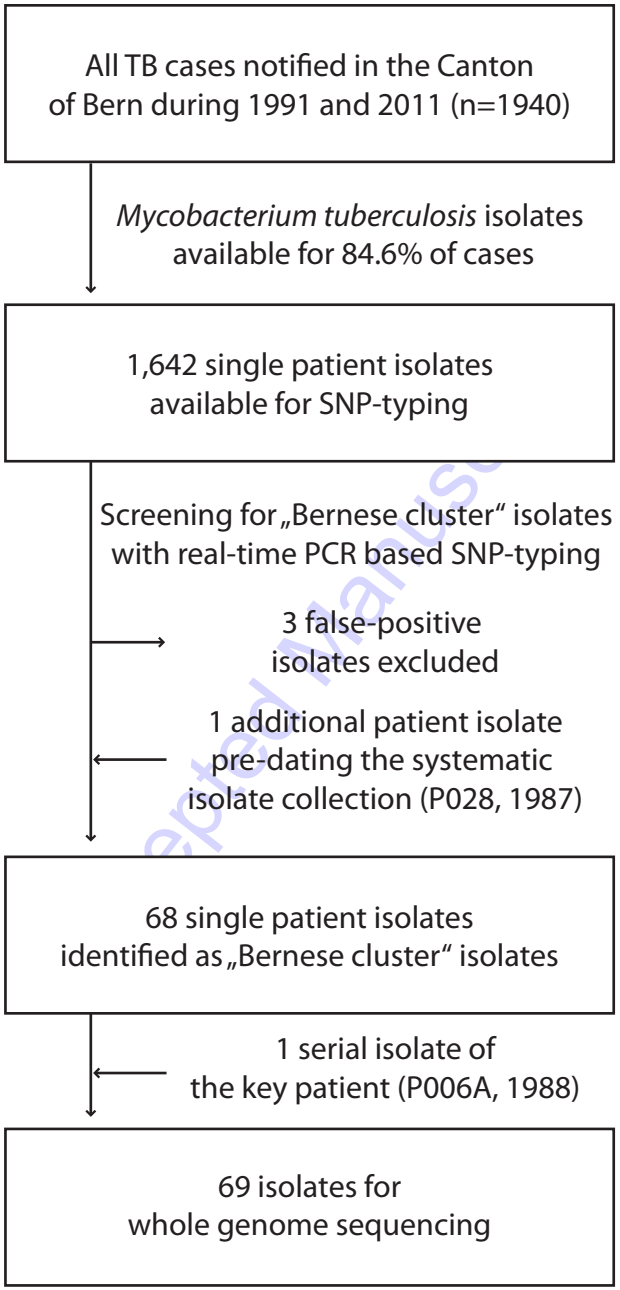
Figure 1. Overview of patient isolates and whole genome sequences generated. A total of 1,642 isolates collected between 1991 and 2011 were available for single nucleotide polymorphism (SNP) genotyping. Three isolates showed ambiguous SNP-typing results and were excluded. One additional patient isolate reported in the original publication [17] and pre-dating the systematic collection of isolates in 1991 was included in the study (P028, isolated in 1987). For the key patient (P006B, isolated in 1991 [17]), a second isolate was available and was included in the genomic analyses.

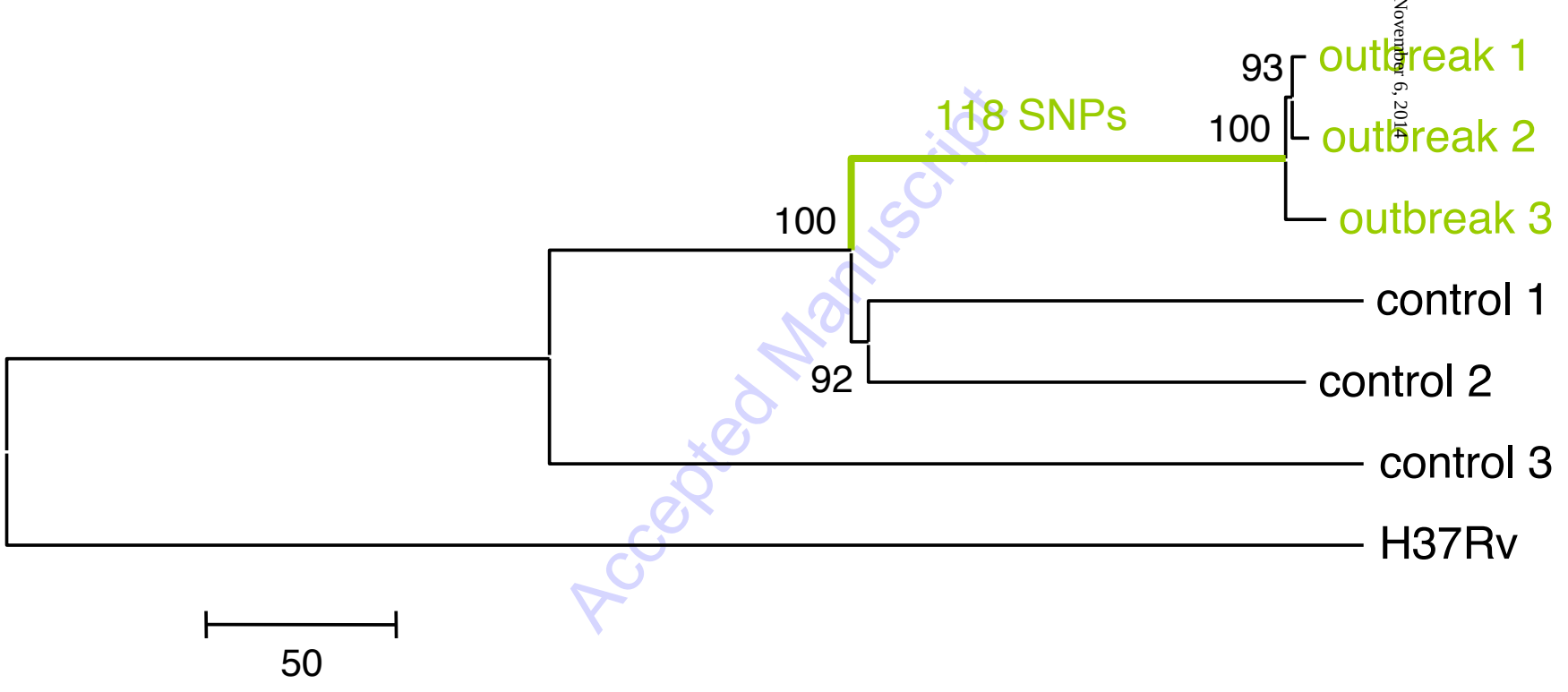
Figure 2. Initial Neighbor Joining phylogeny of *Mycobacterium tuberculosis* isolates. Three whole genomes sequences from the historic outbreak and four control isolates were used to identify single nucleotide polymorphisms specific to the "outbreak" genotype. Node support was assessed by bootstrapping over 1,000 pseudo-replicates and is indicated as percentage.

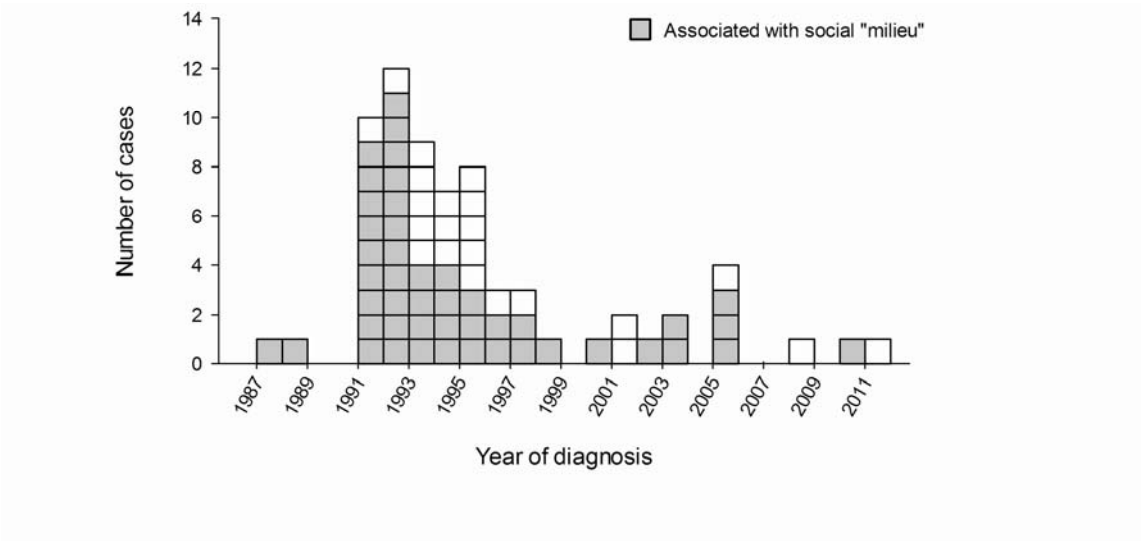
Figure 3. Epidemic curve of the 68 patients identified as tuberculosis cluster cases. Grey boxes indicate patients associated with the social milieu (homeless and/or substance abusers). One additional patient isolate reported in the original publication [17] and pre-dating the systematic collection of isolates in 1991 was included in the study (P028, isolated in 1987). For one patient from 1991 [17], a second isolate was available (P006A, isolated in 1988), and was therefore backdated.

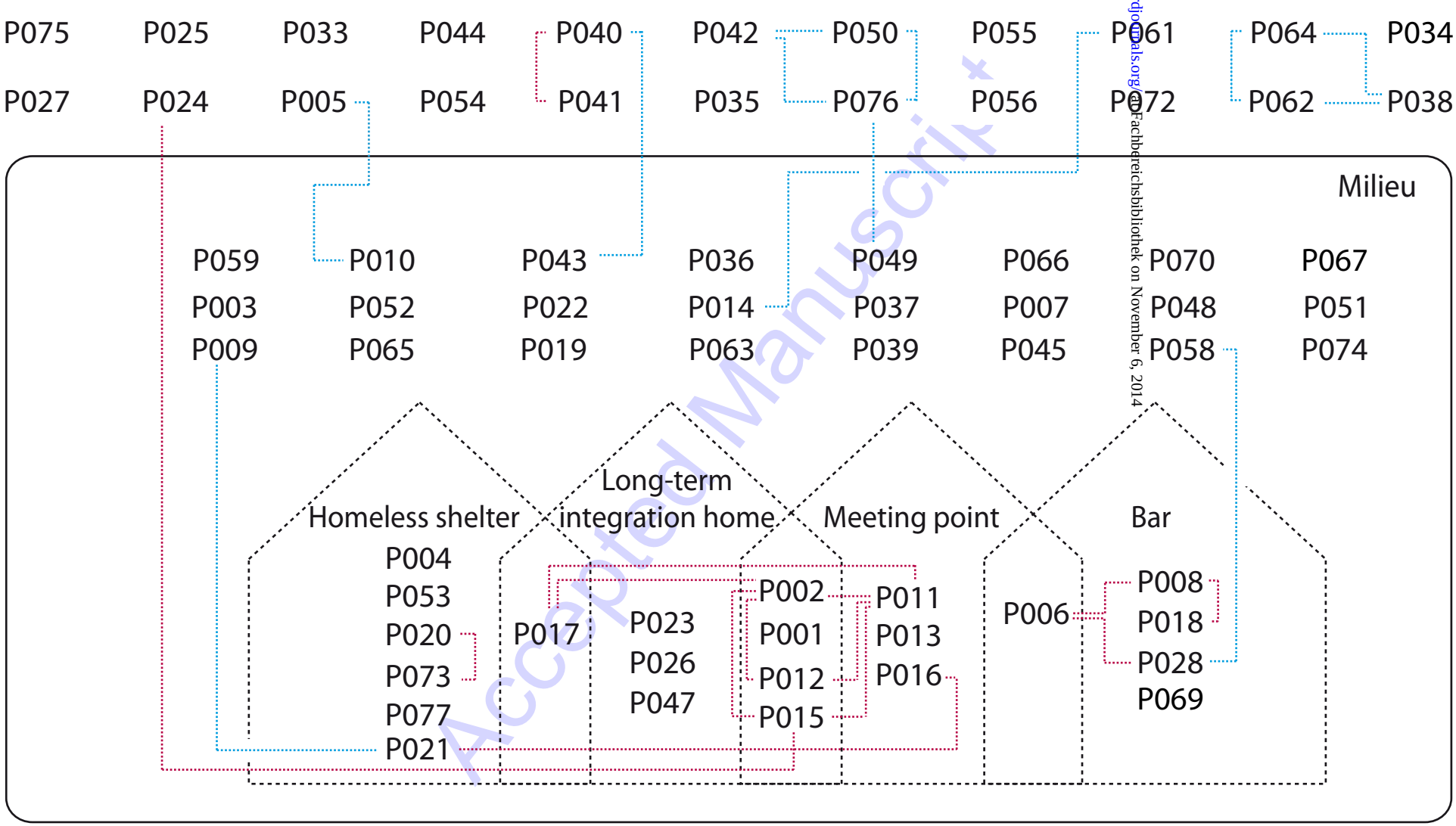
Figure 4. Distribution of tuberculosis (TB) cluster patients in the milieu of substance abusers and homeless people. The four main hotspots of transmission that were identified by social contact tracing are shown (a short-term homeless shelter, a long-term social integration home, a meeting point for substance abusers and a bar). Milieu patients are associated with a particular social context (homeless, substance abuser scene). Red lines indicate confirmed epidemiological links, blue lines suspected social links. Presumptive individual links between milieu patients are not shown, as these patients are highly interlinked. Labels in grey correspond to the more distantly related sub-cluster 4 patients.

Figure 5. Median Joining network using 133 variable single nucleotide positions (SNP) among whole genome sequences of *Mycobacterium tuberculosis* cluster isolates of the "Bernese cluster". Branch lengths correspond to number of SNPs. Circle sizes correspond to number of isolates, Median vectors (mv) are hypothetical genotypes. Position of "mv6" is the root of the network. **A.** Network showing the four identified "sub-clusters" of the cluster, where circle colors correspond to sub-clusters defined. Dark circle colors indicate patients that were associated with the particular social milieu (homeless, substance abuser scene), light circle colors are patients in the non-milieu population. Underlined labels represent isolates that were identified in the original publication [17]. **B.** Network colored according to time period when the *M. tuberculosis* strains were isolated. **C.** Network showing patients associated with a particular hotspot.

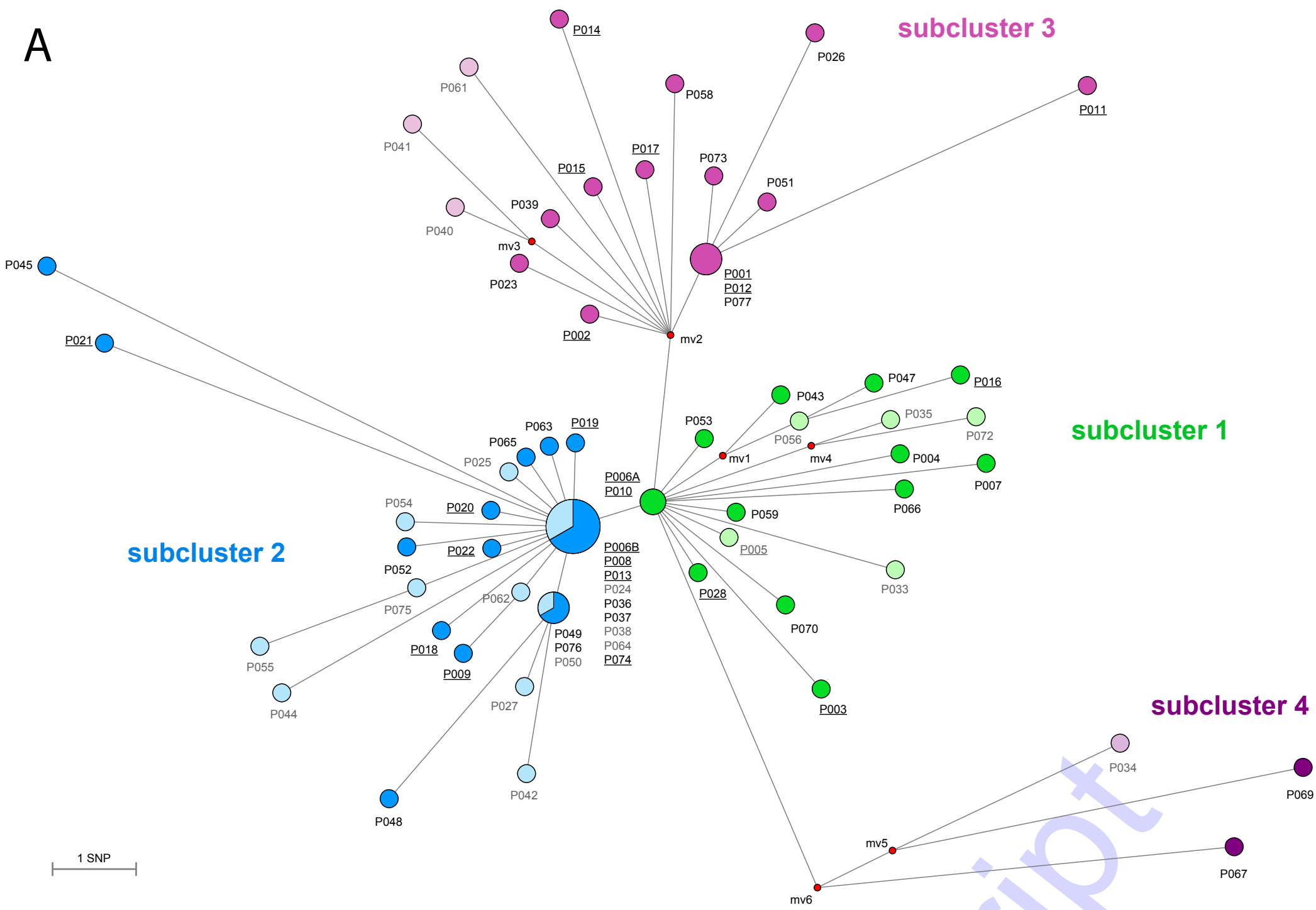

All TB cases notified in the Canton of Bern during 1991 and 2011 (n=1940)
Mycobacterium tuberculosis isolates available for 84.6% of cases
1,642 single patient isolates available for SNP-typing
Screening for „Bernese cluster" isolates with real-time PCR based SNP-typing
3 false-positive isolates excluded
1 additional patient isolate pre-dating the systematic isolate collection (P028, 1987)
68 single patient isolates identified as „Bernese cluster" isolates
1 serial isolate of the key patient (P006A, 1988)
69 isolates for whole genome sequencing



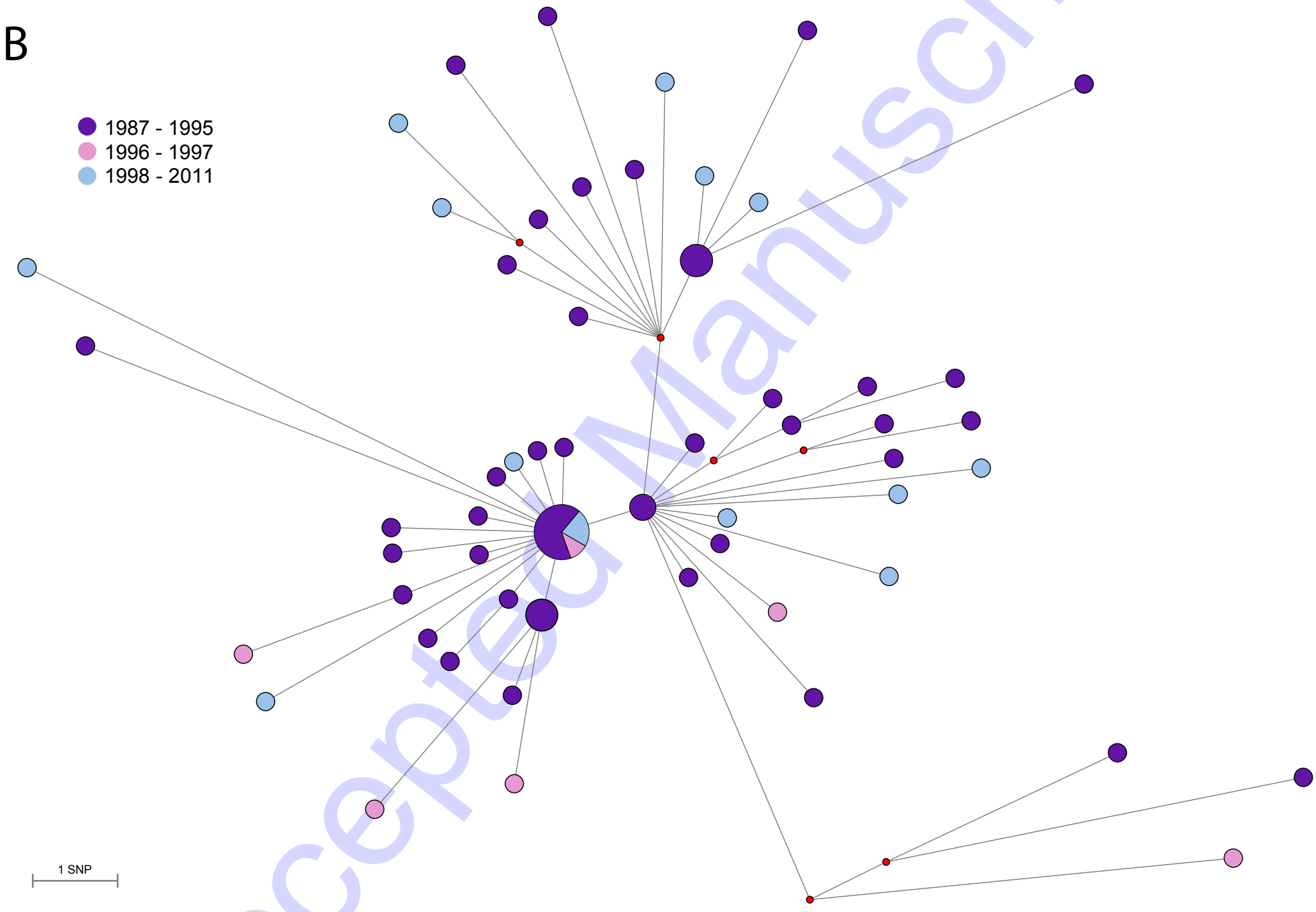




A



B



C